Name: Pearl Law
SCU ID: W0958839
Rank: 3
NMI score: 0.3735

Unsupervised Audio Clustering with Autoencoder and K-means

Introduction

Audio allows us to communicate with each other on many social platforms. There are many audio applications created with deep learning techniques that involve learning to classify sounds and developing predictive audio classification models. The goal of this project is to develop a deep convolutional autoencoder network that extracts meaningful features from a given set of audio samples (via the encoder) and categorizes each sample into one of 20 different categories (via K-means clustering).

Approach

671 audio samples, each of which are 5 seconds long and represent a variety of different sounds (dog, airplane, car horn, etc.), are provided as the training data. Since this is an unsupervised clustering task, no labels are provided as the trained model will predict the label associated with each audio file.

Librosa, a Python package, is used to extract important audio features from the audio samples. To extract key features, the audio files are first converted into spectrograms, which are image representations of sound frequencies changing over time. Mel-frequency cepstral coefficient (MFCC) signals, which describe the overall shape of the sound spectrum, are extracted from each of the audio files. Other than MFCC features, spectral centroid (the center of mass for a sound) and spectral roll off (the frequency below a specified percentage of the total spectral energy) were also tested but generated less desirable results.

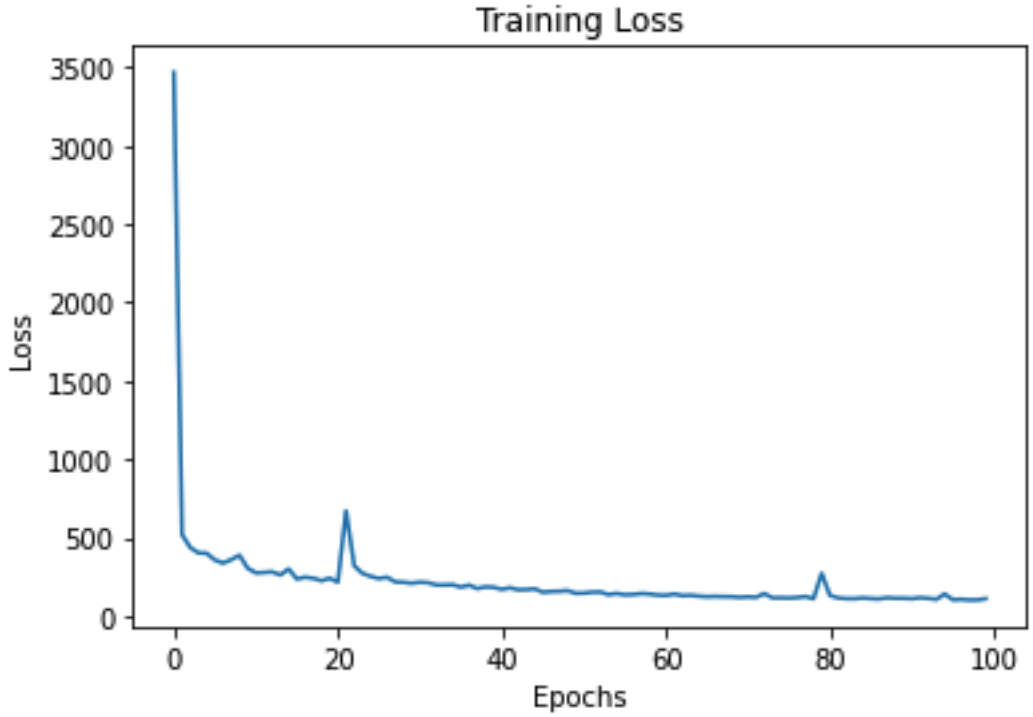
Once features are selected and extracted from each image, each image of size 288 x 432 x 3 is fed into the autoencoder network. The image passes through the encoder, containing four layers of 2D convolution (128, 128, 64, 32 filters, respectively) followed by 2D max pooling layers. The features maps that are extracted get passed to the decoder, containing four layers of 2D convolution (32, 64, 128, 128 filters, respectively) followed by 2D upsampling. The final 2D convolutional layer has 3 filters. All convolutional layers use ReLU activation and have 3 x 3 kernel size. All max pooling layers have 2 x 2 pool size. All layers are same padded to generate an output with the same size as the input. The autoencoder is compiled with Adam optimizer and MSE loss function.

The autoencoder is pre-trained for 100 epochs with a batch size of 4 to obtain the best weights. Early stopping is implemented to stop training if loss remains the same for more than 10 epochs. The best weights are loaded back into the autoencoder and used to build the encoder

weights to make predictions on extracted deep features. The predictions are reshaped, and k-means clustering is performed to determine the label for each audio file.

Results

After pre-training the autoencoder, the loss is minimized to 109.79.



The encoder is trained with the best weights from pre-training the autoencoder and used to make predictions on the audio samples. The normalized mutual information space (NMI) score obtained after k-means clustering is 0.3735. The NMI score obtained evaluates the clustering quality and is equivalent to correctly clustering ~80% of all samples.

Conclusion

Autoencoders are extremely useful for unsupervised learning tasks and efficient at compressing and reconstructing an input from its encoding. In this project, we demonstrated the ability to cluster >80% (or equivalently obtaining an NMI score = 0.3735) on 50% of all audio samples using a 18-layer autoencoder network and k-means clustering algorithm. Future work will focus on optimizing the autoencoder architecture to minimize the loss observed and further improve the clustering quality of audio data.